WHAT IS CLAIMED IS:

1. A method of compressing a log of linguistic data, the log having a plurality of linguistic strings, each string being including at least two tokens, the method comprising:

applying a compression operation to each string;

determining if any two strings match each other after the compression operation; and

removing one of the two matching strings from the log.

2. The method of claim 1, wherein the log is a log of queries.

3. The method of claim 2, wherein the queries are queries relative to a help function.

4. The method of claim 3, wherein the help-related queries are relative to a computer system.

5. The method of claim 1, wherein the compression operation is character-based.

6. The method of claim 1, wherein the compression operation is token-based.

7. The method of claim 1, wherein the compression operation is subsumption.

8. The method of claim 7, wherein subsumption includes applying an impossibility condition to selectively compute edit distance.

9. The method of claim 1, and further comprising:

applying a second compression operation to each string;

determining if any two strings match each other after the second compression operation; and

removing one of the two matching strings from the log.

10. The method of claim 9, wherein the first compression operation is character-based and the second compression operation is token based.

11. The method of claim 10, and further comprising applying subsumption after the second compression operation is complete.

12. The method of claim 11, wherein the subsumption operation is repeated for the log.

13. The method of claim 1, and further comprising training a statistical process with the compressed log.

14. A system for compressing a query log having a plurality of linguistic strings, each string having a plurality of tokens, the system comprising:

an input for receiving a raw query log;

memory for storing the raw query log;

a processor for applying at least one compression operation to each string, and for scanning the modified strings to determine if any match each other so that one of the matching strings can be removed; and

an output for providing a compressed query log once the removal is complete.

15. The system of claim 14, wherein the queries are queries relative to a help function.

16. The system of claim 14, wherein the help-related queries are relative to a computer system.

17. The system of claim 14, wherein the at least one compression operation is character-based.

18. The system of claim 14, wherein the at least one compression operation is token-based.

19. The system of claim 14, wherein the at least one compression operation includes subsumption.

20. The system of claim 19, wherein subsumption includes applying an impossibility condition to selectively compute edit distance.

21. The system of claim 14, and further comprising:

applying at least a second compression operation to each string;

determining if any two strings match each other after the second compression operation; and

removing one of the two matching strings from the log.

22. The system of claim 21, wherein the first compression operation is character-based and the second compression operation is token based.

23. The system of claim 22, and further comprising applying subsumption after the second compression operation is complete.

24. The system of claim 23, wherein the subsumption operation is repeated for the log.

25. The system of claim 14, and further comprising training a statistical process with the compressed log.